arXiv: stat.AP/1502.02406

NONPARAMETRIC BAYESIAN APPROACH TO LR ASSESSMENT IN CASE OF RARE TYPE MATCH

By Giulia Cereda

University of Lausanne
and
Leiden University

The evaluation of a match between the DNA profile of a stain found on a crime scene and that of a suspect (previously identified) involves the use of the unknown parameter $\mathbf{p} = (p_1, p_2, ...)$, (the ordered vector which represents the frequencies of the different DNA profiles in the population of potential donors) and the names of the different DNA types. We propose a Bayesian nonparametric method which models $\mathbf{p}$ through a random variable $\mathbf{P}$ distributed according to the two-parameter Poisson Dirichlet distribution, and discards the information about the names of the different DNA types. The ultimate goal of this model is to evaluate the so-called 'probative value' of DNA matches in the rare type case, that is the situation in which the suspect's profile, matching the crime stain profile, is not in the database of reference.

1. Introduction. The largely accepted method for evaluating how much some available data $\mathcal{D}$ (typically forensic evidence) is helpful in discriminating between two hypotheses of interest (the prosecution hypothesis H_p and the defense hypothesis H_d), is the calculation of the *likelihood ratio* (LR), a statistic that expresses the relative plausibility of the data under these hypotheses, defined as

$$\text{LR} = \frac{\Pr(\mathcal{D}|H_p)}{\Pr(\mathcal{D}|H_d)}. \tag{1}$$

Widely considered the most appropriate framework to report a measure of the 'probative value' of the evidence regarding the two hypotheses (Robertson and Vignaux, 1995; Evett and Weir, 1998; Aitken and Taroni, 2004; Balding, 2005), it indicates the extent to which data is in favor of one hypothesis over the other. Forensic literature presents many approaches to calculate the LR, mostly divided into Bayesian and frequentist methods (see Cereda (2015*b*) for a careful differentiation between these two approaches).

This paper proposes a Bayesian nonparametric method for the LR assessment in the rare type match case, the challenging situation in which there is a match between some characteristic of the recovered material and of the control material, but this characteristic has not been observed yet in previously collected samples (i.e. database of reference). This constitutes a problem because the LR value depends on the proportion of the matching characteristic in a reference population, and this proportion is, in standard practice, estimated using the relative frequency of the characteristic in the available database. In particular, we will focus on Y-STR DNA profile matches, for which the rare type match problem is often recurring (Cereda, 2015*b*). In this case data to evaluate is made of the information about the matching profile and of the list of DNA profile in the database.

The use of a Bayesian nonparametric method is justified by the fact that the parameter of the model is the infinite dimensional vector $\mathbf{p}$, made of the (unknown) sorted population proportions of all possible Y-STR profiles, assumed to be infinitely many. As prior over $\mathbf{p}$ we choose the two parameter Poisson Dirichlet distribution, and treat its parameters as hyperparameters, hence provided with a hyperprior. Moreover, we will discard the information contained in the names of the profiles, and this will lead to a reduction of the data $\mathcal{D}$ to a smaller amount of information D. The reduction of the data can be a wise practice in presence of many nuisance parameters as explained in Cereda (2015*b*), and sometimes the likelihood ratio based on the data reduction is much more precisely estimated than the likelihood ratio based on all data.

The paper is structured in the following way: Section 2 introduces the notation, the assumptions of our model and the prior distribution chosen for parameter $\mathbf{p}$. Section 3 displays the model, via Bayesian network representation, along with some theory on random partitions useful to define a clever and compact representation of the reduced data D. An alternative representation of the same model via the Chinese restaurant process is also described. Section 4 introduces relevant known results regarding the two parameter Poisson Dirichlet distribution, along with a new lemma, which can be used for all the situations in which prosecution and defense agree on the distribution of part of the data and disagree on the distribution of the rest, given the parameter(s). This result will allow to derive the LR in a very elegant way (Section 5). Section 6 displays some experiments of application of this model on a real database of Y-STR profiles, such as model fitting, asymptotic power law behavior, study of the loglikelihood function, and comparison with the LR values obtained in the ideal situation in which vector $\mathbf{p}$ is known. Lastly, Section 7 proposes questions for future research.

2. A Bayesian nonparametric model for the rare type match.

2.1. *The rare type match problem.* In order to evaluate the match between the profile of a particular piece of evidence and a suspect's profile, it is necessary to estimate the proportion of that profile in the population of potential perpetrators. Indeed, it is intuitive that the rarer the matching profile, the more the suspect is in trouble. Problems arise when the observed frequency of the profile in a sample from the population of interest (i.e., in a reference database) is 0. Such characteristic is likely to be rare, but it is challenging to quantify how rare it is. The rare type problem is particularly important in case a new kind of forensic evidence, such as results from DIP-STR markers (see for instance Cereda et al. (2014)) is involved, and for which the available database size is still limited. The same happens when Y-chromosome (or mitochondrial) DNA profiles are used since the set of possible Y-STR profiles is extremely large. As a consequence, most of the Y-STR haplotypes are not represented in the database. The Y-STR marker system will thus be retained here as an extreme but in practice common and important way in which the problem of assessing evidential value of rare type match can arise. This problem is so substantial that it has been defined "the fundamental problem of forensic mathematics" (Brenner, 2010).

The *empirical frequency estimator*, also called *naive estimator*, that uses the frequency of the characteristic in the database, puts unit probability mass on the set of already observed characteristics, and it is thus unprepared for the observation of a new type. A solution could be the *add-constant* estimators (in particular the well known *add-one* estimator, due to Laplace (1814), and the *add-half* estimator of Krichevsky and Trofimov (1981)), which add a constant to the count of each type, included the unseen ones. However, this method requires to know the number of possible unseen types, and it is also not very performing when this number is large compared to the sample size (see Gale and Church (1994) for additional discussion). Alternatively, Good (1953), based on an intuition on A.M. Turing, proposed the *Good Turing estimator* for the total unobserved probability mass, based on the proportion of singleton observations in the sample. An extension of this estimator is applied to the LR assessment in the rare type match in Cereda (2015*b*). For a comparison between *add one* and *Good-Turing* estimator, see Orlitsky et al. (2003). As pointed out in Anevski et al. (2013), the *naive estimator*, and the *Good Turing estimator* are in some sense complementary: the first gives a good estimate for the observed types, and the second for the probability mass of the unobserved ones. More recently, Orlitsky et al. (2004) have introduced the *high profile estimator*, which extends the tail of the *naive estimator* to the region of unobserved types. Anevski et al. (2013) improved this estimator and provided the consistency proof. Papers that address the rare Y-STR haplotype problem in forensic context are for instance Egeland and Salas (2008), Brenner (2010), and Cereda (2015*a*), which applied classical Bayesian approach (the beta binomial and the Dirichlet multinomial problem) to the LR assessment in the rare haplotype case. Moreover, the Discrete Laplace method presented in Andersen et al. (2013), even though not specifically designed for the rare type case can be successfully applied to that extent Cereda (2015*b*).

Bayesian nonparametric estimators for the probability of observing a new type have been proposed by Tiwari and Tripathi (1989), using Dirichlet process, by Lijoi et al. (2007) using general Gibbs prior, and by Favaro et al. (2009) with specific interest to the two parameter Poisson Dirichlet prior. However, for the LR assessment it is required not only the probability of observing a new species but also the probability of observing this same species twice (according to the defense the crime stain profile and the suspect profile are two independent observations), and to our knowledge, this paper is the first one to address the problem of LR assessment in the rare haplotype case using Bayesian nonparametric models. As prior we will use the Poisson Dirichlet distribution, which is proving useful in many discrete domain, in particular language modelling (Teh et al., 2006). In addition, it shows a power law behaviour which describe a incredible variety of phenomena (Newman, 2005).

2.2. *Notation.* Throughout the paper the following notation is chosen: random variables and their values are denoted, respectively, with uppercase and lowercase characters: x is a realization of X. Random vectors and their values are denoted, respectively, by uppercase and lowercase bold characters: $\mathbf{p}$ is a realization of the random vector $\mathbf{P}$. Probability is denoted with $\Pr(\cdot)$, while density of a continuous random variable X is denoted alternatively by $p_X(x)$ or by $p(x)$ when the subset is clear from the context. For a discrete random variable Y, the density notation $p_Y(y)$ and the discrete one $\Pr(Y = y)$ will be alternately used. Moreover, we will use shorthand notation like $p(y \mid x)$ to stand for the probability density of Y with respect to the conditional distribution of Y given $X = x$.

Notice that when using the notation in (1), $\mathcal{D}$ is regarded as events. However, later in the paper, it will be regarded as a random variables. In that case, the following notation will thus be preferred:

$$\text{LR} = \frac{\Pr(\mathcal{D} = d|H_p)}{\Pr(\mathcal{D} = d|H_d)} \quad \text{or} \quad \frac{p(d|H_p)}{p(d|H_d)}. \tag{2}$$

Lastly, notice that “DNA types” is used throughout the paper as a general formula to indicate Y-STR profiles.

2.3. *Model assumptions.* Our model is based on the two following assumptions:

Assumption 1 There are infinitely many DNA types in Nature.

The reason for this assumption is that there are so many possible DNA types that they can be considered infinite. This assumption, already used by e.g. Kimura (1964) in the ‘infinite alleles model', allows to use Bayesian nonparametric methods and avoids the problem of specifying how many different types there are in Nature.

Assumption 2 The names of the different DNA types do not contain information.

The specific sequence of numbers that forms a DNA profile carries information: if two profiles show few differences this means that they are separated by few mutation drifts, hence the profiles share a relatively recent common ancestor. However, this information is difficult to exploit and may be not so relevant for the LR assessment. This is the reason why we will treat DNA types as "colors", and only consider the repartition into different categories. Stated otherwise, we put no topological structure on the space of the DNA types.

Notice that this assumption makes the model a priori suitable for any characteristic which shows many different possible types, thus what written still holds, in principle, also replacing 'DNA types' with any other type. However, we will only test the model with Y-STR data.

2.4. *Prior.* In Bayesian statistics, parameter(s) of interest are modeled through random variables. The (prior) distribution over the parameter(s) should represent the uncertainty about its (their) value(s).

LR assessment for the rare type match involves two unknown parameters of interest: one is $h \in \{H_p, H_d\}$, representing the unknown true hypothesis, the other is $\boldsymbol{p}$, the vector of the unknown population frequencies of all DNA profiles in the population of potential perpetrators. The dichotomous random variable H is used to model parameter h, and the posterior distribution of this random variable, given data, is the ultimate aim of a forensic inquiry. In a similar way, random variable $\boldsymbol{P}$ can be used to model $\boldsymbol{p}$. Because of Assumption 1, $\boldsymbol{p}$ is an infinite dimensional parameter, hence the need of Bayesian nonparametric methods (Hjort et al., 2010). In particular $\boldsymbol{p} = (p_t | t \in T)$, with T a countable set of indexes, $p_t > 0$, and $\sum_t p_t = 1$. Moreover, because of Assumption 2, data will be reduced to random partitions, as explained in Section 3.1, and it will turn out that the distribution of these partitions does not depend on the order of the p_i. Hence, we can force the parameter $\mathbf{p}$ to have values in $\nabla_\infty = \{(p_1, p_2, ...) | p_1 \geq p_2 \geq ..., \sum p_i = 1, p_i > 0\}$, the ordered infinite dimensional simplex. The ordered random vector $\mathbf{p}$ describes an infinite population randomly partitioned into DNA types. The randomness is described by the prior distribution over $\mathbf{p}$, for which we choose the two-parameter Poisson Dirichlet distribution (Pitman and Yor, 1997; Feng, 2010; Buntine and Hutter, 2010; Carlton, 1999; Pitman and Picard, 2006), defined in the following way:

Definition 1 (two parameter GEM distribution). *Given α and θ satisfying the following conditions:*

$$0 \leq \alpha < 1, \ \textit{and } \theta > -\alpha. \tag{3}$$

the vector $\boldsymbol{W} = (W_1, W_2, ...)$ is said to be distributed according to the $\mathrm{GEM}(\alpha, \theta)$, *if*

$$\forall i \quad W_i = V_i \prod_{j=1}^{i-1} (1 - V_j),$$

where V_1, V_2,... are independent random variables distributed according to

$$V_i \sim B(1 - \alpha, \theta + i\alpha).$$

It holds that $W_i > 0$, and $\sum_i W_i = 1$.

The GEM distribution (short for Griffin - Engen - McCloskey distribution') is well known in literature as the "stick breaking prior", since it measures the random sizes in which a stick is broken iteratively. This distribution is invariant under size biased permutation (Engen, 1975), the random permutation defined by sampling from the population and assigning to each type a label, based on the order in which it is first sampled.

Definition 2 (Two parameter Poisson Dirichlet distribution). *Given α and θ satisfying condition (3), and a vector $\boldsymbol{W} = (W_1, W_2, ...) \sim GEM(\alpha, \theta)$, the random vector $\boldsymbol{P} = (P_1, P_2, ...)$ obtained by ordering $\boldsymbol{W}$, such that $p_i \geq p_{i+1}$, is said to be* Poisson Dirichlet distributed *$PD(\alpha, \theta)$. Parameter α is called* discount parameter*, while θ is the* concentration parameter.

Notice that the vector $\boldsymbol{P}$ is obtained by sorting the vector $\boldsymbol{W}$ in non increasing order, while the vector $\boldsymbol{W}$ can be obtained by the so-called *size biased permutation* of the indexes of $\boldsymbol{P}$ (Perman et al., 1992; Pitman and Yor, 1997).

The two parameter Poisson Dirichlet distribution $\mathrm{PD}(\alpha, \theta)$ is the generalization of the well known Poisson Dirichlet distribution with a single parameter θ introduced by Kingman (1975), which is the representation measure (Kingman, 1977, 1978) of the celebrated *Ewens sampling formula* (Ewens, 1972), widely applied in genetics (Karlin and McGregor, 1972; Kingman, 1980). For our model we will not allow $\alpha = 0$, hence we will assume $0 < \alpha < 1$.

It is worth mentioning that an alternative choice for the parameters space is $\alpha < 0$, $\theta = -m\alpha$ for some $m \in \mathbb{N}$ (Pitman, 1996; Gnedin and Pitman, 2006; Gnedin, 2010; Cerquetti, 2010). It corresponds to a model with finitely many (m) DNA types, where the prior over $\mathbf{P} = (P_1, ..., P_m)$ is Dirichlet with m parameters equal to $-\alpha$.

Lastly, we point out that, in practice, we cannot assume to know parameters α and θ: this is why we will treat them as hyperparameters on which we will put an hyperprior.

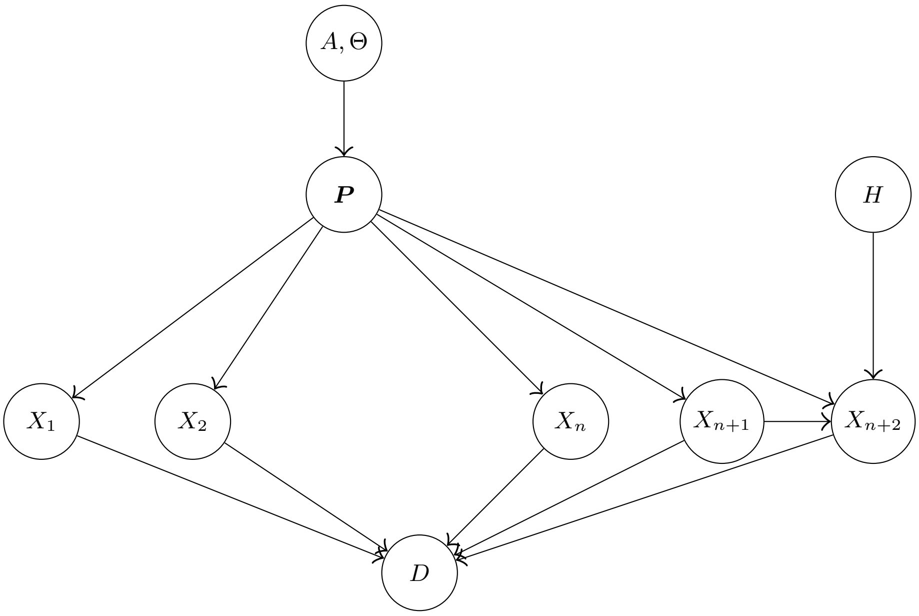


Fig 1. *Bayesian network to show the conditional dependencies of the relevant random variables in our model.*

3. The model. The Bayesian network of Figure 1 encapsulates the conditional dependencies of the variables of the proposed model. They are defined through random variables defined as follows:

- H is a dichotomous random variable that represents the hypotheses of interest and can take values $h \in \{H_p, H_d\}$, according to the prosecution or the defense, respectively. A uniform prior on the hypotheses is chosen:
$$\Pr(H = h) \propto 1 \quad \text{for } h = \{H_p, H_d\}.$$
- (A, Θ) is the random vector that represents the hyperparameters α and θ, satisfying condition (3). The joint distribution of these two parameters (hyperprior) will be generically denoted as $p(\alpha, \theta)$:
$$(A, \Theta) \sim p(\alpha, \theta).$$
- The random vector $\boldsymbol{P}$ with values in ∇_∞, represents the ranked population frequencies. $\boldsymbol{P} = \boldsymbol{p} = (p_1, p_2, ...)$ means that p_1 is the frequency of the most common DNA type in the population, p_2 is the frequency of the second most common DNA type, and so on. As a prior for $\boldsymbol{P}$ we use the two-parameter Poisson Dirichlet distribution as in Definition 2:
$$\boldsymbol{P}|(A, \Theta) = (\alpha, \theta) \sim PD(\alpha, \theta).$$
- Integer valued random variables X_1, ..., X_n represent the ranks of the population proportions of the DNA types of the individuals in the database (after some arbitrary ordering for profiles in the database is chosen). For instance, $X_3 = 5$ means that the third individual in the database has the fifth most common DNA type in the population. Since $\boldsymbol{p}$ is unknown these random variables cannot be observed. Given $\boldsymbol{p}$ they are an i.i.d. sample from $\boldsymbol{p}$:
$$X_1, X_2, ..., X_n|\boldsymbol{P} = \boldsymbol{p} \sim_{i.i.d.} \boldsymbol{p}. \tag{4}$$
- X_{n+1} represents the rank of the suspect's DNA type. It is again a draw from $\boldsymbol{p}$.
$$X_{n+1}|\boldsymbol{P} = \boldsymbol{p} \sim \boldsymbol{p}.$$
- X_{n+2} represents the rank of the crime stain's DNA type. According to the prosecution, given X_{n+1}, this random variable is deterministic (it is equal to x_{n+1} with probability 1). According to the defense it is another sample from $\boldsymbol{p}$:
$$X_{n+2}|\boldsymbol{P} = \boldsymbol{p}, X_{n+1} = x_{n+1}, H = h \sim \begin{cases} \delta_{x_{n+1}} & \text{if } h = H_p \\ \boldsymbol{p} & \text{if } h = H_d \end{cases}.$$

As already mentioned, $X_1, ..., X_{n+2}$ can not be observed. They represent the database ranked according to the unknown rank in $\boldsymbol{p}$ and constitute an intermediate layer that helps in expressing the data in terms of observable partitions. Section 3.1 recalls some notions about random partitions, useful before defining node D, representing the 'reduced' data we want to evaluate.

3.1. *Random partitions.* A *partition of a set* A is an unordered collection of nonempty and disjoint subsets of A the union of which forms A. Particularly interesting for our model are partitions of the set $A = [n] = \{1, ..., n\}$, denoted as $\pi_{[n]}$. The set of all partitions of $[n]$ will be denoted as $\mathcal{P}_{[n]}$. Random partitions of $[n]$ will be denoted as $\Pi_{[n]}$. In addition, a *partition of* n is a finite non increasing sequence of positive integers that sum up to n. Partitions of n will be denoted as π_n.

Given a sequence of integer valued random variables $X_1, ..., X_n$, let $\Pi_{[n]}(X_1, X_2, ..., X_n)$ be the random partition defined by the equivalence classes of their indexes using the random equivalence relation $i \sim j$ iff $X_i = X_j$. This construction allows to build a map from the set of values of $X_1, ..., X_n$ to the set of the partitions of $[n]$ as in the following example ($n = 10$):

$$\begin{aligned} \mathbb{N}^{10} &\to \mathcal{P}_{[10]} \\ X_1, ..., X_{10} &\longmapsto \Pi_{[10]}(X_1, X_2, ..., X_{10}) \\ (3, 1, 3, 1, 2, 2, 6, 9, 4, 1) &\longmapsto \{\{1, 3\}, \{2, 4, 10\}, \{5, 6\}, \{7\}, \{8\}, \{9\}\} \end{aligned}$$

Typical data to evaluate in case of a match is $\mathcal{D} = (E, B)$, where $E = (E_s, E_t)$, and

- B = the database of size n, which contains a sample of DNA types, indexed by $i = 1, ..., n$, from the population of possible perpetrators,
- E_s = suspect's DNA type,
- E_t = crime stain's DNA type (matching with the suspect's type).

In agreement with Assumption 2, we can consider the reduction of data which ignores information about the names of the DNA types: this is achieved, for instance, by retaining only the equivalence classes of the relation "to have the same DNA type".

The database can thus be reduced to the partition of $[n]$, denoted $\pi^{\text{Db}}_{[n]}$, and obtained using the equivalence classes of the indexes. Notice that the same partition is obtained via random variables $X_1, ..., X_n$, as defined in (4). Stated otherwise, we can reduce B to $\pi^{\text{Db}}_{[n]}$, the partition of $[n]$ obtained from the database. However, data is actually made of the background data along with the evidence, two new observations that match. In a similar way, when the suspect profile is considered we obtain the partition $\pi^{\text{Db+}}_{[n+1]}$, where the first n integers are partitioned as in $\pi^{\text{Db}}_{[n]}$, and $n + 1$ constitutes a single subset (at least in the rare type match case).

When the crime stain profile is considered we obtain the partition $\pi^{\text{Db++}}_{[n+2]}$ where the first n integers are partitioned as in $\pi^{\text{Db}}_{[n]}$, and $n + 1$ and $n + 2$ belongs to the same (new) subset.

Random variables $\Pi^{\text{Db}}_{[n]}$, $\Pi^{\text{Db+}}_{[n+1]}$, and $\Pi^{\text{Db++}}_{[n+2]}$ are used to model $\pi^{\text{Db}}_{[n]}$, $\pi^{\text{Db+}}_{[n+1]}$, and $\pi^{\text{Db++}}_{[n+2]}$, respectively.

Notice that, given α and θ, prosecution and defense agree on the distribution of $\Pi^{\text{Db+}}_{[n+1]}$ but disagree on the distribution of $\Pi^{\text{Db++}}_{[n+2]}$.

It is worth noticing that, by construction, the same random partitions can be defined through random variables X_1, ..., X_{n+2}:

$$\begin{aligned} \Pi^{\text{Db}}_{[n]} &= \Pi_{[n]}(X_1, ..., X_n), \\ \Pi^{\text{Db+}}_{[n+1]} &= \Pi_{[n+1]}(X_1, ..., X_{n+1}), \\ \Pi^{\text{Db++}}_{[n+2]} &= \Pi_{[n+2]}(X_1, ..., X_{n+2}). \end{aligned}$$

To clarify, consider the following example of a database (Db) with $k = 6$ different DNA types, from $n = 10$ individuals:

$$\text{Db} = (h_1, h_2, h_1, h_2, h_3, h_3, h_4, h_5, h_6, h_2),$$

where h_i is the name of the ith DNA type in the order chosen for the database. This can be reduced to the partition of $[10]$:

$$\pi^{\text{Db}}_{[10]} = \{\{1,3\},\{2,4,10\},\{5,6\},\{7\},\{8\},\{9\}\}.$$

Then, the part of data prosecution and defense agree on is

$$\pi^{\text{Db+}}_{[11]} = \{\{1,3\},\{2,4,10\},\{5,6\},\{7\},\{8\},\{9\},\{11\}\},$$

while the entire data D can be represented as

$$\pi^{\text{Db++}}_{[12]} = \{\{1,3\},\{2,4,10\},\{5,6\},\{7\},\{8\},\{9\},\{11,12\}\}.$$

Now, assume that $\boldsymbol{p}$ is known, thus we know also that h_1 is, for instance, the fourth most frequent type, h_2 is the second most frequent type, and so on. Stated otherwise, we are now able to observe the variables X_1, ..., X_{n+2}: $X_1 = 4$, $X_2 = 2$, $X_3 = 4$, $X_4 = 2$, $X_5 = 3$, $X_6 = 3$, $X_7 = 10$, $X_8 = 13$, $X_9 = 5$, $X_{10} = 2$, $X_{11} = 9$, $X_{12} = 9$. It is easy to check that $\Pi_{[n]}(X_1, ..., X_n) = \pi^{\text{Db}}_{[n]}$, $\Pi_{[n+1]}(X_1, ..., X_{n+1}) = \pi^{\text{Db+}}_{[n+1]}$, $\Pi_{[n+2]}(X_1, ..., X_{n+2}) = \pi^{\text{Db++}}_{[n+2]}$.
Data D can now be defined as:

- $D = \pi^{\text{Db++}}_{[n+2]}$, the partition of $[n+2]$ obtained from the database enlarged with the two new observations.

Node D of Figure 1 is defined accordingly. Notice that, given $X_1, ..., X_{n+2}$, D is deterministic. A very relevant result is that, according to Proposition 4 in Pitman (1992) it is possible to describe directly the distribution of $D \mid \alpha, \theta, H$. Hence, we can get rid of the intermediate layer of nodes X_1, ..., X_{n+2}. In particular, it holds that if

$$\boldsymbol{P} \mid \alpha, \theta \sim PD(\alpha, \theta),$$

and

$$X_1, X_2, ... \mid \boldsymbol{P} = \boldsymbol{p} \sim_{\text{i.i.d}} \boldsymbol{p},$$

then, for all n, the random partition $\Pi_{[n]} = \Pi_{[n]}(X_1, ..., X_n)$ has the following distribution:

$$\mathbb{P}^{\alpha,\theta}_n(\pi_{[n]}) := \Pr(\Pi_{[n]} = \pi_{[n]} | \alpha, \theta) = \frac{[\theta+\alpha]_{k-1;\alpha}}{[\theta+1]_{n-1;1}} \prod_{i=1}^{k} [1-\alpha]_{n_i-1;1}, \tag{5}$$

where n_i is the size of the ith block of $\pi_{[n]}$ (the blocks are here ordered according to the least element), and $\forall x, b \in \mathbb{R}, a \in \mathbb{N}$, $[x]_{a,b} := \begin{cases} \prod_{i=1}^{a-1}(x+ib) & \text{if } a \in \mathbb{N}\backslash\{0\} \\ 0 & \text{if } a = 0 \end{cases}$. This formula is also known as the *Pitman sampling formula*, further studied in Pitman (1995). The model of Figure 1 can thus be simplified (see Figure 2).

It holds that $\Pr(D|\alpha, \theta, H_p) = \mathbb{P}^{\alpha,\theta}_{n+1}(\pi^{\text{Db+}}_{[n+1]})$, and $\Pr(D|\alpha, \theta, H_d) = \mathbb{P}^{\alpha,\theta}_{n+2}(\pi^{\text{Db++}}_{[n+2]})$.
Notice that for $\alpha = 0$ we obtain the Ewens's sampling formula.

3.2. *Chinese Restaurant representation.* There is an alternative characterization of this model, called "Chinese restaurant process", due to Aldous (1985) for the one parameter case, and studied in details for the two parameter version in Pitman and Picard (2006). It is defined as follows: consider a restaurant with infinite tables, each one infinitely large. Let $Y_1, Y_2, ...$ be integer valued random variables that represent the seating plan: tables

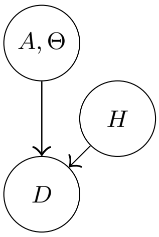


FIG 2. *Simplified version of the Bayesian network in Figure 1*

are ranked in order of occupancy, and $Y_i = j$ means that the ith customer seats at the jth table to be created. The process is described by the following transition matrix:

$$Y_1 = 1$$

$$\Pr(Y_{n+1} = i | Y_1, ..., Y_n) = \begin{cases} \dfrac{\theta + k\alpha}{n+\theta} & \text{if } i = k+1 \\ \\ \dfrac{n_i - \alpha}{n+\theta} & \text{if } 1 \leq i \leq k \end{cases} \tag{6}$$

where k is the number of tables occupied by the first n customers, and n_i is the number of customers that occupy table i.

$Y_1, ..., Y_n$ are not i.i.d., nor exchangeable, but it holds that $\Pi_{[n]}(Y_1, ..., Y_n)$ has the same distribution as $\Pi_{[n]}(X_1, ..., X_n)$, with $X_1, ..., X_n$ defined as in (4) (in particular they are distributed according to the Pitman sampling formula (5)).

Stated otherwise, we can use the seating plan of n customers to obtain the same partition $\pi^{\text{Db}}_{[n]}$ built through the database (or by partitioning $X_1, ..., X_n$). Then $\pi^{\text{Db+}}_{[n+1]}$ is obtained when a new customer has chosen an unoccupied table (remember we are in the rare type case), and $\pi^{\text{Db++}}_{[n+2]}$ is obtained when the $n+2$nd customer goes to the same table of the $n+1$st (suspect and crime stain have the same DNA type). In particular, thanks to (**??**), we can write

$$p(\pi^{\text{Db++}}_{[n+2]} \mid H_p, \pi^{\text{Db+}}_{[n+1]}, \alpha, \theta) = 1, \tag{7}$$

and

$$p(\pi^{\text{Db++}}_{[n+2]} \mid H_d, \pi^{\text{Db+}}_{[n+1]}, \alpha, \theta) = \frac{1-\alpha}{n+1+\theta} \tag{8}$$

since the $n+2$nd customer goes to the same table of the $n+1$st where he seats alone.

4. Some results.

This section presents some useful results that will be used in the forthcoming sections. In particular, Lemma 4.1, suitable to broader applications, is here applied to simplify the LR development. Then, some results from Pitman and Picard (2006) regarding the two parameter Poisson Dirichlet distribution, are listed.

4.1. *A useful Lemma.*

The following lemma is a result regarding four general random variables A, X, Y, H whose conditional dependencies are described by the Bayesian network of Figure 4.1. The importance of this result is due to the possibility of applying it to a very common forensic situation: the prosecution and the defense disagree on the entirety of data (Y) but agree on a part of it (X) (indeed, as already noticed, defense and prosecution agree on the distribution of $\pi^{\text{Db+}}_{[n+1]}$, but not on the distribution of $\pi^{Db++}_{[n+2]}$). Data depends on parameters (A).

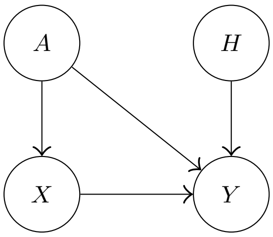


FIG 3. *Conditional dependencies of the random variables of Lemma4.1*

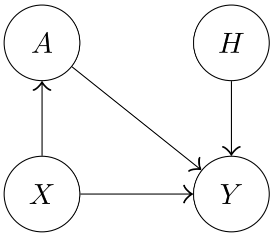
LEMMA 4.1. *Given four random variables A, H, X and Y, whose conditional dependencies are represented by the Bayesian network of Figure 4.1, the likelihood function for h, given $X = x$ and $Y = y$ satisfies*

$$\mathrm{lik}(h \mid x, y) \ \propto \mathbb{E}(p(y \mid x, A, h) \mid X = x).$$

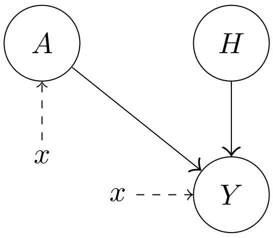
PROOF. The model of Figure 4.1 represents four variables A, H, X and Y whose joint probabilty density can be factored as

$$p(a, h, x, y) \ = \ p(a)\, p(x \mid a)\, p(h)\, p(y \mid x, a, h).$$

By Bayes formula, $p(a)\, p(x \mid a) = p(x)\, p(a \mid x)$. This rewriting corresponds to reversing the direction of the arrow between A and X:



The random variable X is now a root node. This means that when we probabilistically condition on $X = x$, the graphical model changes in a simple way: we can delete the node X, but just insert the value x as a parameter in the conditional probability tables of the variables A and Y which formerly had an arrow from node X. The next graph represents this model:



This tells us, that conditional on $X = x$, the joint density of A, Y and H is equal to

$$p(a \mid x) p(h) p(y \mid x, a, h).$$

The joint density of H and Y is obtained by integrating out the variable a. It can be expressed as a conditional expectation value, since $p(a \mid x)$ is the density of A given $X = x$. We find:

$$p(h)\mathrm{E}(p(y \mid x, A, h) \mid X = x).$$

Recall that this is the joint density of two of our variables, H and Y, after conditioning on the value $X = x$. Let us now also condition on $Y = y$. It follows that the density of H given $X = x$ and $Y = y$ is proportional (as function of H, for fixed x and y) to the same expression, $p(h)\mathrm{E}(p(y \mid x, A, h) \mid X = x)$.

This is a product of the prior for h with some function of x and y. Since posterior odds equals prior odds times likelihood ratio, it follows that the likelihood function for h, given $X = x$ and $Y = y$ satisfies

$$\mathrm{lik}(h \mid x, y) \ \propto \mathrm{E}(p(y \mid x, A, h) \mid X = x).$$

□

Corollary 4.1. *Given four random variables A, H, X and Y, whose conditional dependencies are represented by the network of Figure 4.1, the likelihood ratio for $H = h_1$ against $H = h_2$ given $X = x$ and $Y = y$ satisfies*

$$LR = \frac{\mathbb{E}(p(y|x, A, h_1)|X = x)}{\mathbb{E}(p(y|x, A, h_2)|X = x)}. \tag{9}$$

4.2. *Known results about the two parameter Poisson Dirichlet distribution.* We will now list some theoretical results which will be useful in the forthcoming analysis. Most of these results can be found in Pitman and Picard (2006).

Denote as K_n the random number of blocks of a partition $\Pi_{[n]}$ distributed according to the Pitman sampling formula with parameters α and θ.

- It exists a positive random variable S_α such that

  $$\lim_{n \to +\infty} \frac{K_n}{n^\alpha} = S_\alpha \quad \text{a.s.} \tag{10}$$

  the distribution of S_α is a generalization of the Mittag Leffler distribution (Gorenflo et al., 2014).
- If $\boldsymbol{P} \sim \mathrm{PD}(\alpha, \theta)$, then

  $$P_i \to Z i^{-1/\alpha}, \quad \text{a.s., when } i \to +\infty \tag{11}$$

  for a random variable Z such that $Z^{-\alpha} = \Gamma(1-\alpha)/S_\alpha$.
- For a fixed $\alpha \in (0, 1)$, the $\mathrm{PD}(\alpha, \theta)$ (for different θ) are all mutually absolutely continuous. This means that θ cannot be consistently estimated for α in the range of interest. On the other hand, the power law behavior described above tells us that α can be consistently estimated.
- Studying (6) one can see that when n increases, the parameter θ becomes less and less important. However, it describes how much “social” are the customers: the smaller θ the more the customers tend to seat to already occupied tables. Thus, it determines the sizes of the big tables, but it won’t be much important for our application (the more rare DNA types).

- Given Π_n distributed according to Pitman sampling formula (5), it holds that

$$\lim_{n\to+\infty}\frac{m_j(n)}{n^\alpha}=\frac{\alpha\Gamma(j-\alpha)}{\Gamma(1-\alpha)j!}S_\alpha \quad \text{a.s.}\,\forall j \tag{12}$$

  where $m_j(n)$, $j=1,...,n$ the random number of blocks of the partition $\Pi_{[n]}$ of size j. This result is presented in Gnedin et al. (2007), based on Karlin (1967).

5. The likelihood ratio.

The hypotheses of interest are:

- H_p = The crime stain was left by the suspect.
- H_d = The crime stain was left by someone else.

The LR will thus be defined as

$$\text{LR}=\frac{p(\pi^{\text{Db++}}_{[n+2]}|H_p)}{p(\pi^{\text{Db++}}_{[n+2]}|H_d)}=\frac{p(\pi^{\text{Db+}}_{[n+1]},\pi^{\text{Db++}}_{[n+2]}|H_p)}{p(\pi^{\text{Db+}}_{[n+1]},\pi^{\text{Db++}}_{[n+2]}|H_d)}.$$

where the last equality is due to the fact that $\Pi^{\text{Db+}}_{[n+1]}$ is deterministic, given $\Pi^{\text{Db++}}_{[n+2]}$.

Corollary 4.1 with $A=(A,\Theta)$, $X=\Pi^{\text{Db+}}_{[n+1]}$, $Y=\Pi^{\text{Db++}}_{[n+2]}$, and $H=H$ allows to obtain the LR as:

$$\begin{aligned}\text{LR} &= \frac{\mathbb{E}(p(\pi^{\text{Db++}}_{[n+2]}\mid\pi^{\text{Db+}}_{[n+1]},A,\Theta,H_p)\mid\Pi^{\text{Db+}}_{[n+1]}=\pi^{\text{Db+}}_{[n+1]})}{\mathbb{E}(p(\pi^{\text{Db++}}_{[n+2]}\mid\pi^{\text{Db+}}_{[n+1]},A,\Theta,H_d)\mid\Pi^{\text{Db+}}_{[n+1]}=\pi^{\text{Db+}}_{[n+1]})}\\ &=\frac{1}{\mathbb{E}\Big(\frac{1-A}{n+1+\Theta}\mid\Pi^{\text{Db+}}_{[n+1]}=\pi^{\text{Db+}}_{[n+1]}\Big)}.\end{aligned}$$

where the last equality is due to (7) and (8). By defining the random variable $\Phi=n\dfrac{1-A}{n+1+\Theta}$ we can write the LR as

$$\text{LR}=\frac{n}{\mathbb{E}(\Phi\mid\Pi_{[n+1]}=\pi_{[n+1]})}. \tag{13}$$

5.1. *True LR.*

In order to evaluate the performance of this method one would like to compare the LR values obtained with (13) with the 'true' ones, meaning the LR values obtained when vector $\mathbf{p}$ is known, which means that we have the list of the frequencies of all the DNA types in the population of interest. The LR in this case can be obtained in the following way:

$$\text{LR}=\frac{p(\pi^{\text{Db++}}_{[n+2]}|\pi^{\text{Db+}}_{[n+1]},H_p,\mathbf{p})}{p(\pi^{\text{Db++}}_{[n+2]}|\pi^{\text{Db+}}_{[n+1]},H_d,\mathbf{p})}=\frac{1}{p(\pi^{\text{Db++}}_{[n+2]}|\pi^{\text{Db+}}_{[n+1]},H_d,\mathbf{p})} \tag{14}$$

$$=\frac{1}{\Pr(X_{n+2}=X_{n+1}|\pi^{\text{Db+}}_{[n+1]},H_d,\mathbf{p})} \tag{15}$$

$$=\frac{1}{\sum\limits_{(x_1,...,x_{n+1})}\Pr(X_{n+2}=x_{n+1}|x_1,...,x_{n+1},\pi^{\text{Db+}}_{[n+1]},H_d,\mathbf{p})p(x_1,...,x_{n+1}|\pi^{\text{Db+}}_{[n+1]},\mathbf{p})} \tag{16}$$

$$=\frac{1}{\sum\limits_{(x_1,...,x_{n+1})}p_{x_{n+1}}p(x_1,...,x_{n+1}|\pi^{\text{Db+}}_{[n+1]},\mathbf{p})} \tag{17}$$

$$=\frac{1}{\mathbb{E}(p_{x_{n+1}}|\pi^{\text{Db+}}_{[n+1]},\mathbf{p})}. \tag{18}$$

Notice that in the rare type case x_{n+1} is observed only once among the x_1, ..., x_{n+1}. Hence we call it a singleton. Let N_1 denote the number of singletons, and $\mathcal{S}$ the set of all singletons. Given $\mathbf{p}$ and $\pi^{\text{Db+}}_{[n+1]}$, it holds that the distribution of X_{n+1} is the same as the distribution of all other N_1 singletons. This implies that:

$$N_1\mathbb{E}(p_{x_{n+1}}|\pi^{\text{Db+}}_{[n+1]},\mathbf{p})=\mathbb{E}(\sum_{x_i\in\mathcal{S}}p_{x_i}|\pi^{\text{Db+}}_{[n+1]},\mathbf{p}).$$

Let us denote as X_1^*, .., X_K^* the K different values taken by X_1, ..., X_{n+1}, ordered according to the frequency of their values. Stated otherwise, if n_i is the frequency of x_i^* among $x_1,...,x_{n+1}$, then $n_1\geq n_2\geq ...\geq n_K$. Moreover, in case X_i^* and X_j^* have the same frequency ($n_i=n_j$), than they are ordered according to their values. For instance, if $X_1=4$, $X_2=2$, $X_3=4$, $X_4=2$, $X_5=3$, $X_6=3$, $X_7=10$, $X_8=13$, $X_9=5$, $X_{10}=2$, $X_{11}=9$, then $X_1^*=2, X_2^*=3, X_3^*=4, X_4^*=5, X_5^*=9, X_6^*=10, X_7^*=13$.

By definition, it holds that

$$\mathbb{E}(\sum_{x_i\in\mathcal{S}}p_{x_i}|\pi^{\text{Db+}}_{[n+1]},\mathbf{p})=\mathbb{E}(\sum_{j\colon n_j=1}p_{x_j^*}|\pi^{\text{Db+}}_{[n+1]},\mathbf{p}).$$

Notice that $(n_1,n_2,...,n_K)$ is a partition of $n+1$, which will be denoted as $\pi^{\text{Db+}}_{n+1}$. In the example, $\pi^{\text{Db+}}_{n+1}=(3,2,2,1,1,1,1)$. A more compact representation for $\pi^{\text{Db+}}_{n+1}$ can be obtained by using two vectors $\mathbf{a}$ and $\mathbf{r}$ where a_j are the distinct numbers occurring in the partition, ordered, and each r_j is the number of repetitions of a_j. J is the length of these two vectors, and it holds that $n+1=\sum_{j=1}^J a_jr_j$. In the example above we have that $\pi^{\text{Db+}}_{n+1}=(\mathbf{a},\mathbf{r})$ with $\mathbf{a}=(1,2,3)$ and $\mathbf{r}=(4,2,1)$.

Since the distribution of $\sum\limits_{j\colon n_j=1}p_{x_j^*}$ only depends on $\pi^{\text{Db+}}_{n+1}$, the latter can replace $\pi^{\text{Db+}}_{[n+1]}$. Thus, it holds that

$$\text{LR}=\frac{N_1}{\mathbb{E}(\sum\limits_{j\colon n_j=1}p_{x_j^*}|\pi^{\text{Db+}}_{n+1},\mathbf{p})}. \tag{19}$$

Notice that the knowledge of $\mathbf{p}$, is not enough to observe $X_1^*,...,X_K^*$: $\mathbf{p}$ is sorted in decreasing order, and even if we know the different values p_i, we don't know to which category each value belongs. There is a function, χ, treated here as latent variable, which assigns all DNA types, ordered according to their frequency in Nature, to one of the number $\{1,2,...,J\}$ corresponding to the position in $\mathbf{a}$ of its frequency in the sample, or to 0 if the type if not observed. Stated otherwise,

$$\chi:\{1,2,...\}\longrightarrow\{1,2,...,J\}.$$

$$\chi(i)=\begin{cases}0 & \text{if the } i\text{th most common species in Nature is not observed in the sample,}\\ j & \text{if the } i\text{th most common species in Nature is one of the } r_j \text{ observed } a_j \text{ times in the sample.}\end{cases}$$

Given $\pi^{\text{Db+}}_{n+1}=(\mathbf{a},\mathbf{r})$, χ must satisfy the following conditions:

$$\sum_{i=1}^{\infty}\mathbf{1}_{\chi(i)=j}=r_j,\qquad\forall j. \tag{20}$$

The map χ can be represented by a vector $\boldsymbol{\chi}=(\chi_1,\chi_2,...)$ made of its values: $\chi_i=\chi(i)$. In the example above we have that $\chi=(0,3,2,2,1,0,0,0,1,1,0,0,1,0...0)$.

Notice that, given $\pi_{n+1}^{\mathrm{Db+}} = (\mathbf{a}, \mathbf{r})$, the knowledge of $\boldsymbol{\chi}$ implies the knowledge of X_1^*, ..., X_K^*: indeed it is enough to sort the positive values among the χ_i and take their positions in χ solving ties by considering the position themselves (if $\chi_i = \chi_j$, than the order is given by i and j). For instance, in the example, if we sort the values of χ and we collect their positions we get $(2, 3, 4, 5, 9, 10, 13)$: the reader can notice that we got back to the $X_1^*, ..., X_7^*$.

This means that to obtain the distribution of $X_1^*, ..., X_K^* | \pi_{n+1}^{\mathrm{Db+}}, \mathbf{p}$, which appears in (19), it is enough to obtain the distribution of $\boldsymbol{\chi} | \pi_{n+1}^{\mathrm{Db+}}, \mathbf{p}$. Actually, we are only interested in the mean of the sum of singletons in samples of size $n+1$ from the distribution of $X_1^*, ..., X_K^* | \pi_{n+1}^{\mathrm{Db+}}, \mathbf{p}$: this means that we can just simulate samples from the distribution of $\boldsymbol{\chi} | \pi_{n+1}, \mathbf{p}$ and sum those p_a such that $\chi_a = 1$.

To simulate samples from the distribution of $\boldsymbol{\chi} | \pi_{n+1}, \mathbf{p}$ we use a Metropolis - Hasting algorithm, on the space of the vectors satisfying condition (20). Notice that for the model we assumed $\mathbf{p}$ to be infinitely long, but for simulations we will use a finite $\bar{\mathbf{p}}$, of length M. This is equivalent to assume that only M elements in the infinite $\mathbf{p}$ are positive, and the remaining infinite tail is made of zeros. Then the state space of the Metropolis Hasting Markov chain is made of all vectors of length M whose elements belong to $\{0, 1, ..., J\}$, and satisfy the condition (20). If we start with a initial point $\boldsymbol{\chi}_0$ which satisfies (20) and, at each allowed move of the Metropolish Hasthing , we swap two different values χ_a and χ_b inside the vector, condition (20) remains satisfied. The algorithm is based on a similar one proposed in Anevski et al. (2013).

This method allows us to obtain the 'true' LR when the vector $\mathbf{p}$ is known. This is rarely the case, but we can put ourselves in a fictitious world where we know $\mathbf{p}$, and compare the true values for the LR with the one obtained by applying our model when $\mathbf{p}$ is unknown. This will be done in the forthcoming section.

6. Analysis on a real database.

In this section we present the study we made on a database of 18,925 Y-STR 23-loci profiles from 129 different locations in 51 countries in Europe (Purps et al., 2014) [1]. Different analyses are performed by considering only 7 Y-STR loci (DYS19, DYS389 I, DYS389 II, DYS3904, DYS3915, DY3926,DY3937) but similar results have been observed with the use of 10 loci.

First the maximum likelihood estimators α_{MLE} and θ_{MLE} using the entire database are obtained. Their values are $\alpha_{\mathrm{MLE}} = 0.5$ and $\theta_{\mathrm{MLE}} = 216$.

In order to check if the choice of the two parameter Poisson Dirichlet prior is a sensible one we first compare the ranked frequencies from the database with the relative frequencies of several samples of size n obtained from realisations of $\mathrm{PD}(\alpha_{MLE}, \theta_{MLE})$. The asymptotic behaviour described in (11) is also checked. Lastly, we will analyse the loglikelihood function for data $\pi_{[n+1]}^{\mathrm{Db+}}$ in order to study the denominator of the LR.

6.1. *Model fitting.*

In Figure 4, the ranked frequencies from the database are compared to the relative frequencies of samples of size n obtained from several realizations of $\mathrm{PD}(\alpha_{MLE}, \theta_{MLE})$. To do so we run several times the Chinese Restaurant seating plan (up to $n = 18,925$ customers): each run is equivalent to generate a new realization $\boldsymbol{p}$ from the $\mathrm{PD}(\alpha_{MLE}, \theta_{MLE})$. The partition of the customers into tables is the same as the partition obtained from an i.i.d. sample of size n from $\boldsymbol{p}$. The ranked relative sizes of each table (thin lines) are compared to the ranked frequencies of our database (thick line).

[1] The database has previously been cleaned by Mikkel Meyer Andersen (`http://people.math.aau.dk/~mikl/?p=y23`).

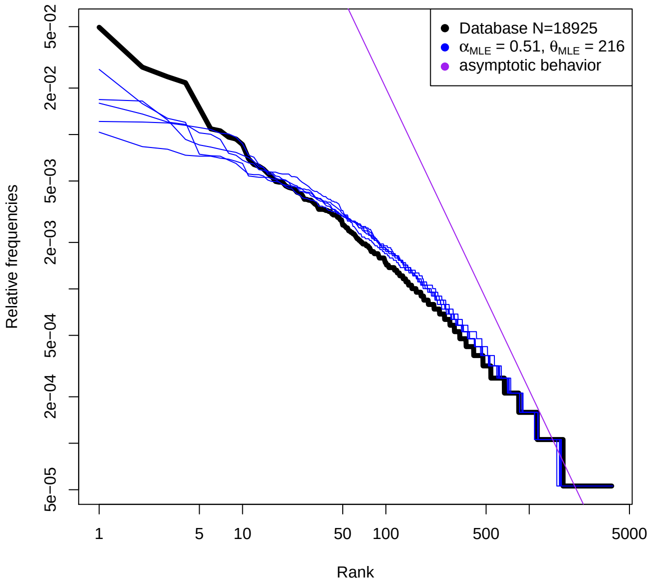


FIG 4. *Log scale ranked frequencies from the database (thick line) are compared to the relative frequencies of samples of size n obtained from realization of PD$(\alpha_{MLE}, \theta_{MLE})$ (thin lines). Asymptotic power law behavior is also displayed (dotted lines).*

6.2. *Asymptotic power law behavior.* The asymptotic behavior described in (11) is also analyzed in Figure 4. This behavior is expected in the tail (the limit is over i) and clearly the number of customers ($n = 18,925$) is not big enough for the small P_i to follow the power law.

6.3. *Loglikelihood.* It is also interesting to investigate the shape of the loglikelihood function for α and θ given $\pi^{\text{Db++}}_{[n+1]}$. It is defined as

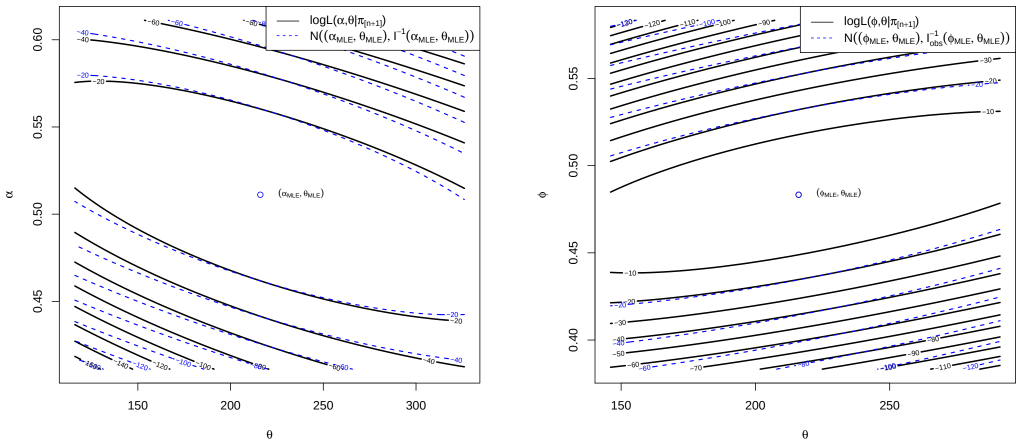
$$l_{n+1}(\alpha, \theta) := \log p(\pi^{\text{Db++}}_{[n+1]} | \alpha, \theta).$$

In Figure 5 (a), the loglikelihood function is compared to the Gaussian distribution centered in the maximum likelihood estimates for α and θ, with the observed Fisher information as covariance matrix. In Figure 5 (b) the loglikelihood reparametrized using $\phi = n\dfrac{1-\alpha}{n+1+\theta}$, and θ instead of α and θ, is displayed and compared to the corresponding Gaussian distribution.

The posterior distribution for (Φ,Θ) given $\Pi^{\text{Db+}}_{[n+1]}$ is proportional to the loglikelihood $l_{[n+1]}(\phi, \theta)$ times the prior $p(\phi, \theta)$. The Gaussian behavior of $l_{[n+1]}(\phi, \theta)$ is particularly interesting since it allows to conclude that if the prior $p(\phi, \theta)$ is smooth around $(\phi_{MLE}, \theta_{MLE})$, we can approximate $\mathbb{E}(\Phi | \Pi^{\text{Db++}}_{[n+1]})$ with $\phi_{MLE} = n\dfrac{1-\alpha_{MLE}}{n+1+\theta_{MLE}}$. Hence, one can approximate the LR itself in the following way:

$$\text{LR} \approx \frac{n+1+\theta_{MLE}}{1-\alpha_{MLE}}. \tag{21}$$

Notice that this is equivalent to an hybrid approach, in which the parameters are estimated through the MLE (frequentist) and their value plugged into the Bayesian LR.



(a) Relative loglikelihood for parameters α and θ, compared to a Gaussian distribution, 95% and 99% confidence intervals (green and red)

(b) Relative loglikelihood for $\phi = n\frac{1-\alpha}{n+1+\theta}$ and θ compared to a Gaussian distribution 95% and 99% confidence intervals (green and red).

Fig 5.

6.4. *Analyzing the error.* As explained in Section 5.1, a Metropolis Hasting algorithm, based on Anevski et al. (2013), can be used to obtain the 'true LR', that is the LR when the vector $\mathbf{p}$ is known, as defined in (14) - (18). The latter will be denoted as $\mathrm{LR}_{|\mathbf{p}}$. This can be compared to the LR obtained with the method described in this paper, when $\mathbf{p}$ is unknown, as defined in (5) - (13). Notice that errors appear at different levels (Cereda, 2015*b*): error due to limitedness of samples, error in the model, error in the choice of the parameters of the model. The following three tests will explore these levels.

Test 1. Instead of using the big database of Purps et al. (2014) we consider its restriction to the Dutch population (of size 2085): we pretend this to be the entire population of possible perpetrators, and compare the distribution of $\log_{10}(\mathrm{LR}_{|\mathbf{p}})$ and $\log_{10}\mathrm{LR}$ obtained by 100 samples of size 100 from this population.

Test 2. In order to avoid error due to model selection, we use as entire population a sample from a realization from $\mathrm{PD}(\alpha,\theta)$ distribution. This is done by running the two parameter Chinese restaurant process up to 2085 customers. Then, again, 100 samples of size 100 from this population are sampled and $\log_{10}(\mathrm{LR}_{|\mathbf{p}})$ and $\log_{10}\mathrm{LR}$ are compared. This procedure is repeated 5 times to use 5 different Poisson Dirichlet populations.

Test 3. The same as in Test 2, but in order to avoid also error due to MLE parameter estimations, we use as parameters α and θ for $\log_{10}\mathrm{LR}$ exactly those which have been used to generate the Chinese restaurant process.

7. Future research questions.

Because of the mutual absolute continuity results we know that θ cannot be consistently estimated. However, there exists at least one consistent estimator for α (Carlton, 1999), namely:

$$\hat{\alpha} = \frac{\log K_n}{\log n}.$$

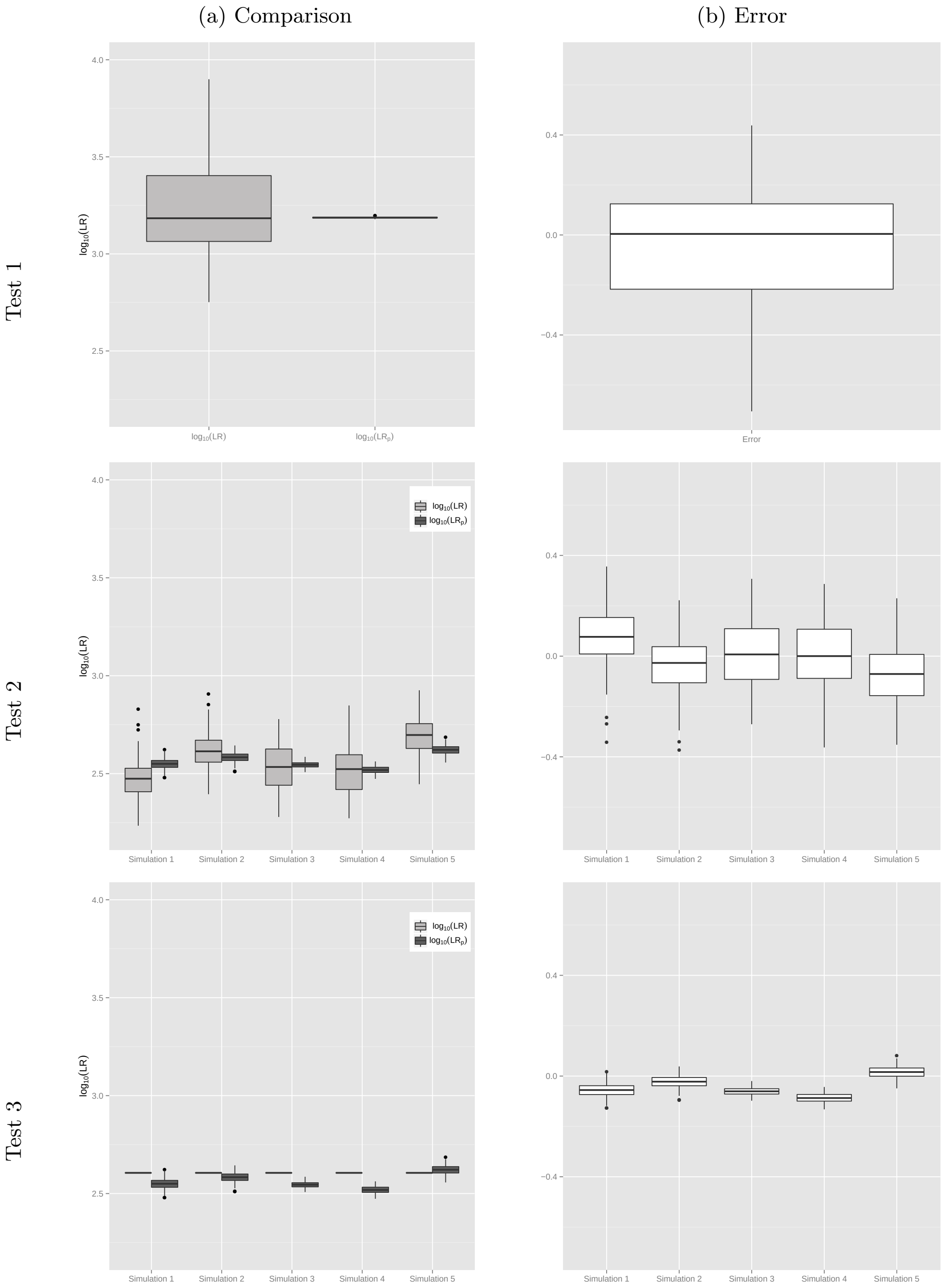


Fig 6. *For Test 1, Test 2 and Test 3 we plot: (a) comparison between the distribution of* $\log_{10} LR$ *when the population proportions* $\mathbf{p}$ *are known (*$LR_{|\mathbf{p}}$*), and when they are not (*LR*). (b) the error* $\log_{10} LR_{|\mathbf{p}} - \log_{10} LR$.

Moreover, α can be estimated consistently also from the power law of (11). We are interested in the consistency of α_{MLE}, at least when θ is known, although literature (Carlton, 1999) is quite skeptical about its performance. The observed Fisher information for α grows with n and this gives some hope for the consistency of α_{MLE}.

The Gaussian behavior of Figure 5 was unexpected. At least, we expect that increasing n, α and θ would become independent, thus the ellipses will rotate.

Acknowledgment. I am indebted to Jim Pitman and Alexander Gnedin for their help in understanding their important theoretical results, and to Mikkel Meyer Andersen for providing a cleaned version of the database of Purps et al. (2014). This research was supported by the Swiss National Science Foundation, through grants no. 105311-144557 and 10531A-156146, and carried out in the context of a joint research project, supervised by Franco Taroni (University of Lausanne, Ecole des sciences criminelles), and Richard Gill (Mathematical Institute, Leiden University).

References.

Aitken, C. and Taroni, F. (2004), *Statistics and the Evaluation of Evidence for Forensics Scientists*, John Wiley & Sons, Chichester.

Aldous, D. J. (1985), *Exchangeability and related topics*, Vol. 1117 of *École D'Été de Probabilités de Saint-Flour*, Springer-Verlag.

Andersen, M. M., Eriksen, P. S. and Morling, N. (2013), 'The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies', *Journal of Theoretical Biology* **329**, 39–51.

Anevski, D., Gill, R. D. and Zohren, S. (2013), 'Estimating a probability mass function with unknown labels', http://arxiv.org/abs/1312.1200.

Balding, D. (2005), *Weight-of-evidence for Forensic DNA Profiles*, John Wiley & Sons Hoboken, NJ.

Brenner, C. H. (2010), 'Fundamental problem of forensic mathematics—The evidential value of a rare haplotype', *Forensic Science International: Genetics* **4**, 281–291.

Buntine, W. and Hutter, M. (2010), 'A Bayesian View of the Poisson-Dirichlet Process', http://arxiv.org/abs/1007.0296.

Carlton, M. A. (1999), Applications of the Two-Parameter Poisson-Dirichlet Distribution, PhD thesis, University of California, Los Angeles.

Cereda, G. (2015*a*), 'Impact of model choice on LR assessment in case of rare haplotype match (bayesian approach)', arXiv:1502.02406.

Cereda, G. (2015*b*), 'Impact of model choice on LR assessment in case of rare haplotype match (frequentist approach)', arXiv:1502.04083.

Cereda, G., Biedermann, A., Hall, D. and Taroni, F. (2014), 'An investigation of the potential of DIP-STR markers for DNA mixture analyses', *Forensic Science International: Genetics* **11**, 229 – 240.

Cerquetti, A. (2010), 'Bayesian nonparametric analysis for a species sampling model with finitely many types', http://arxiv.org/abs/1001.0245.

Egeland, T. and Salas, A. (2008), 'Estimating haplotype frequency and coverage of databases', *PLoS ONE* **3**, e3988–e3988.

Engen, S. (1975), 'A note on the geometric series as a species frequency model', *Biometrika* **62**, 694–699.

Evett, I. and Weir, B. (1998), *Interpreting DNA evidence: Statistical Genetics for Forensic Scientists*, Sinauer Associates, Sunderland.

Ewens, W. (1972), 'Sampling theory of selectively neutral alleles', *Theoretical Population Biology* **3**(1), 87–112.

Favaro, S., Lijoi, A., Mena, R. H. and Pruenster, I. (2009), 'Bayesian nonparametric inference for species variety with a two parameter poisson-dirichlet process prior', *Journal of the Royal Statistical Society: Series B (Methodological)* **71**, 993–1008.

Feng, S. (2010), *The Poisson-Dirichlet Distribution and Related Topics: Models and Asymptotic Behaviors*, Springer.

Gale, W. A. and Church, K. W. (1994), What's wrong with adding one?, *in* 'Corpus-Based Research into Language. Rodolpi'.

Gnedin, A. (2010), 'A Species Sampling Model with Finitely many Types', *Electronic Communications in Probability* **15**, 79–88.

Gnedin, A., Hansen, B. and Pitman, J. (2007), 'Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws', *Probability Surveys* pp. 146–171.

Gnedin, A. and Pitman, J. (2006), 'Exchangeable gibbs partitions and stirling triangles', *Journal of Mathematical Sciences* **138**, 5674–5685.

Good, I. (1953), 'The population frequencies of species and the estimation of population parameters', *Biometrika* **40**, 237–264.

Gorenflo, R., Kilbas, A., Mainardi, F. and Rogosin, S. (2014), *Mittag-Leffler Functions, Related Topics and Applications*, Springer Monographs in Mathematics, Springer Berlin Heidelberg.

Hjort, N., Holmes, C., Müller, P. and Walker, S. (2010), *Bayesian Nonparametrics*, Cambridge University Press.

Karlin, S. (1967), 'Central limit theorems for certain infinite urn schemes', *Journal of Mathematics and Mechanics* **17**, 373–401.

Karlin, S. and McGregor, J. (1972), 'Addendum to a paper of w. ewens', *Theoretical population biology* **3**, 113–116.

Kimura, M. (1964), 'The number of alleles that can be maintained in a finite population', *Genetics* **49**, 725–738.

Kingman, J. (1977), 'The population structure associated with the ewens sampling formula', *Theoretical Population Biology* **11**, 274 – 283.

Kingman, J. (1978), 'Random partitions in population-genetics', *Proceedings of the Royal Society of London series A-Mathematical physical and Engineering sciences* **361**, 1–20.

Kingman, J. (1980), *Mathematics of Genetic Diversity*, Society for Industrial and Applied Mathematics.

Kingman, J. F. C. (1975), 'Random discrete distributions', *Journal of the Royal Statistical Society. Series B (Methodological)* **37**, 1–22.

Krichevsky, R. and Trofimov, V. (1981), 'The performance of universal coding', *IEEE Transactions on Information Theory* **27**, 199–207.

Laplace, P. (1814), *Essai philosophique sur les probabilites*, M.me V.e Courcier.

Lijoi, A., Mena, R. H. and Pruenster, I. (2007), 'Bayesian nonparametric estimation of the probability of discovering new species', *Biometrikaiom* **94**, 769–786.

Newman, M. (2005), 'Power laws, Pareto distributions and Zipf's law', *Contemporary Physics* **46**, 323–351.

Orlitsky, A., Santhanam, N. P., Viswanathan, K. and Zhang, J. (2004), On Modeling Profiles Instead of Values, *in* 'UAI', pp. 426–435.

Orlitsky, A., Santhanam, N. and Zhang, J. (2003), 'Always Good Turing: asymptotically optimal probability estimation', *Science (New York, N.Y.)* **302**, 427–431.

Perman, M., Pitman, J. and Yor, M. (1992), 'Size-biased sampling of poisson point-processes and excursions', *Probability Theory and Related Fields* **92**, 21–39.

Pitman, J. (1992), The two-parameter generalization of ewens' random partition structure, Technical report 345, Department of Statistics U.C. Berkeley CA.

Pitman, J. (1995), 'Exchangeable and partially exchangeable random partitions', *Probability Theory and Related Fields* **102**, 145–158.

Pitman, J. (1996), Some Developments of the Blackwell-MacQueen Urn Scheme, *in* 'Statistics, Probability and Game Theory; Papers in honor of David Blackwell', pp. 245–267.

Pitman, J. and Picard, J. (2006), *Combinatorial Stochastic Processes*, Combinatorial Stochastic Processes: École D'Été de Probabilités de Saint-Flour XXXII - 2002, Springer.

Pitman, J. and Yor, M. (1997), 'The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator', *Annals of Probability* **25**, 855–900.

Purps, J., Siegert, S., Willuweit, S., Nagy, M., Alves, C., Salazar, R., Angustia, S. M. T., Santos, L. H., Anslinger, K., Bayer, B., Ayub, Q., Wei, W., Xue, Y., Tyler-Smith, C., Bafalluy, M. B., Martínez-Jarreta, B., Egyed, B., Balitzki, B., Tschumi, S., Ballard, D., Court, D. S., Barrantes, X., Bäßler, G., Wiest, T., Berger, B., Niederstätter, H., Parson, W., Davis, C., Budowle, B., Burri, H., Borer, U., Koller, C., Carvalho, E. F., Domingues, P. M., Chamoun, W. T., Coble, M. D., Hill, C. R., Corach, D., Caputo, M., D'Amato, M. E., Davison, S., Decorte, R., Larmuseau, M. H. D., Ottoni, C., Rickards, O., Lu, D., Jiang, C., Dobosz, T., Jonkisz, A., Frank, W. E., Furac, I., Gehrig, C., Castella, V., Grskovic, B., Haas, C., Wobst, J., Hadzic, G., Drobnic, K., Honda, K., Hou, Y., Zhou, D., Li, Y., Hu, S., Chen, S., Immel, U.-D., Lessig, R., Jakovski, Z., Ilievska, T., Klann, A. E., García, C. C., de Knijff, P., Kraaijenbrink, T., Kondili, A., Miniati, P., Vouropoulou, M., Kovacevic, L., Marjanovic, D., Lindner, I., Mansour, I., Al-Azem, M., Andari, A. E., Marino, M., Furfuro, S., Locarno, L., Martín, P., Luque, G. M., Alonso, A., Miranda, L. S., Moreira, H., Mizuno, N., Iwashima, Y., Neto, R. S. M., Nogueira, T. L. S., Silva, R., Nastainczyk-Wulf, M., Edelmann, J., Kohl, M., Nie, S., Wang, X., Cheng, B., Núñez, C., Pancorbo, M. M. d., Olofsson, J. K., Morling, N., Onofri, V., Tagliabracci, A., Pamjav, H., Volgyi, A., Barany, G., Pawlowski, R., Maciejewska, A., Pelotti, S., Pepinski, W., Abreu-Glowacka, M., Phillips, C., Cárdenas, J., Rey-Gonzalez, D., Salas, A., Brisighelli, F., Capelli, C., Toscanini, U., Piccinini, A., Piglionica, M., Baldassarra, S. L., Ploski, R., Konarzewska, M., Jastrzebska, E., Robino, C., Sajantila, A., Palo, J. U., Guevara, E., Salvador, J., Ungria, M. C. D., Rodriguez, J. J. R., Schmidt, U., Schlauderer, N., Saukko, P., Schneider, P. M., Sirker, M., Shin, K.-J., Oh, Y. N., Skitsa, I., Ampati, A., Smith, T.-G., Calvit, L. S. d., Stenzl, V., Capal, T., Tillmar, A., Nilsson, H., Turrina, S., De Leo, D., Verzeletti, A., Cortellini, V., Wetton, J. H., Gwynne, G. M., Jobling, M. A., Whittle, M. R., Sumita, D. R., Wolańska-Nowak, P., Yong, R. Y. Y., Krawczak, M., Nothnagel, M. and Roewer, L. (2014), 'A global analysis of y-chromosomal haplotype diversity for 23 str loci', *Forensic Science International: Genetics* **12**, 12–23.

Robertson, B. and Vignaux, G. A. (1995), *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*, John Wiley & Sons, Chichester.

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006), 'Hierarchical dirichlet processes', *Journal of the American Statistical Association* **101**, 1566–1581.

Tiwari, R. C. and Tripathi, R. C. (1989), 'Nonparametric bayes estimation of the probability of discovering a new species.', *Communications in statistics: Theory and methods* **A18**, 877–895.

University of Lausanne, Ecole des sciences criminelles,
Faculté de droit, des sciences criminelles et d'administration publique
1015, Lausanne-Dorigny, Switzerland,
E-mail: giulia.cereda@unil.ch